Moral Realism, Naturalism, and Moral Disagreement

Billy Dunaway
University of Missouri–St Louis

Draft of 31st March 2022

Moral realism is a metaphysical thesis about the nature of morality. The realist holds that morality is a deep and significant part of the world. It is real, and hence a part of the world, in the way the subatomic particles studied by the physicist are, or the planets and solar systems studied by the astrophysicist. Moreover the reality of morality, according to the realist, consists in the fact that it is a deep part of the world, in the sense that it explains other phenomena. The subatomic particles of physics explain the behavior of macrophysical objects. The movements of planetary bodies explain the short duration of daylight during winter in the northern hemisphere. Morality concerns properties such as moral rightness and obligation. Should these be a deep part of reality, then they explain something as well: candidates include the appropriateness of blame, praise, and the presence (or absence) of moral motivations in agents.

There are several views about the nature of morality that are incompatible with moral realism. Emotivists and expressivists hold that moral sentences ('stealing is wrong') primarily express an emotion or a plan for action (disapproval of stealing ((Blackburn 1988); a plan not to steal (Gibbard 2003)). This is primarily a claim about the function of moral sentences, and is on its face silent about whether moral obligation is a deep part of reality. But expressivists have argued that expressive function of moral language serves to explain the features of morality that the realist gives a metaphysical explanation for. Expressivists and emotivists thus typically reject the central claims of realism.

Subjectivists hold that moral facts are in the world, but locate them in a place that makes subjectivism incompatible with realism. For the subjectivist, what makes stealing wrong for me is a fact in the world, namely that I disapprove of stealing. The sentence 'stealing is wrong' as uttered by me says something about the world, for the subjectivist, since it says something about my attitude toward stealing. The subject-matter of moral sentences is, however, not a deep or significantly explanatory one; 'stealing is wrong' as uttered by Hulk Hogan says something about a different set of attitudes, namely Hulk Hogan's. The subjectivist view is also incompatible with realism.

Finally, views such as error theory and fictionalism hold that there are no moral truths at all. These views agree with the realist that moral claims aim to state truths about the world. They deny, however, that there are any such truths to be found. The error theorist thus holds that moral sentences such as 'stealing is wrong' are systematically false, because there is no attribute of wrongness that stealing instantiates. Fictionalists, while they agree with this claim about reality, hold that moral sentences should be understood not to be making (false) claims about reality, but instead are to be interpreted as making potentially true claims about a fictional world. According to the fictionalist, just as 'Holmes lives at 223b Baker Street' is true in the fictional world of Sherlock Holmes, so also 'stealing is wrong' is true according to the fiction of morality.

There are several versions of moral realism. The most well-known division distinguishes between naturalist and non-naturalist moral realism. Naturalists hold that, while moral obligation is a deep part of reality, it is also a natural property, in some sense. What makes a property natural is a matter for debate, but one common characterization is that moral obligation is natural if and only if it is identical with a property that can be studied by an empirical science. Other formulations of naturalism use the ideology of grounding, holding that moral properties are grounded in natural properties. Non-naturalism, by contrast, is a version of realism that denies the central naturalist thesis. Non-naturalists hold that obligation is not identical to, or grounded by, any natural property. Instead, obligation is a "sui generis" property, distinct in kind from natural properties. It is a distinctively moral property.

Opponents of moral realism object to the view on a number of grounds. One objection is that realism is ontologically extravagant, as it takes on commitments about the reality of moral properties which are unnecessary. There are other accounts of the nature of morality (according to the objection) which do just as well without the metaphysical commitments of realism. This objection requires that expressivism, subjectivism, or some other competitor to moral realism can be adequately spelled out.

Another objection is that accounting for morality in terms of the reality of morality is bound to be inadequate. According to this line of thought, morality is essentially practical, in the sense that it bears on what we should do. Morality is, in philosophical jargon, normative. Facts about reality, however, are facts about how the world is, and so are not practical; the fact that the world is a certain way always leaves open what we should do about it. Reality cannot account for the normativity of morality.

A third objection to moral realism raises the phenomenon of moral disagreement. I will spell this objection out in more detail in section 1, but the rough outlines of the objection are as follows: people who use moral sentences are capable of disagreeing with each other. Some English speakers sincerely assert the sentence 'capital punishment is wrong' and others sincerely assert 'capital punishment is permissible'. In asserting these sentences, speakers *disagree* with each other. The disagreement is of the same type as a disagreement in which one English speaker sincerely asserts 'the Earth is flat' and another asserts 'the Earth is round'.

Objectors claim that the phenomenon of disagreement is incompatible with moral realism. Disagreements about factual matters—such as a disagreement over whether the Earth is flat—can be explained by widespread agreement surrounding the topic that is at issue. Flat-earthers and their opponents both agree on what flatness and roundness consist in, and on what it would take for the Earth to be flat (or not). The sentences they assent to express this disagreement concern the same subject-matter, and express conflicting claims about what it is like. Morality, on the other hand, is different: there is no substantial agreement about whether morality is about reducing suffering and promoting pleasure, following universal prescriptions of reason, or something else. If morality is real, as the moral realist holds, then at most one of these views manages to accurately describe the nature of moral reality. Others fail to capture it, and describe something else (or nothing at all). Proponents of these alternative views are thus not talking about moral reality at all. The realist must hold that what appear to be moral disagreements are, implausibly, instances of speakers *talking past one another*.

This chapter will focus on objections from disagreement in more detail. §1 sketches some well-known instances of the objection in the literature. §2 raises the relevance of naturalist vs. non-naturalist versions of moral realism. §3 finishes by outlining some prominent lines of response on behalf of the realist.

1 Disagreement in the literature

A common objection to moral realism in the literature is that it cannot explain moral disagreement. Or, more specifically, it cannot explain the *extent* of moral disagreement. R. M. Hare gives an early instance of this kind of argument:

> Let us suppose that a missionary, armed with a grammar book, lands on a cannibal island. The vocabulary of his grammar book gives him the equivalent, in the cannibals' language, of the English word 'good'. Let us suppose that, by a strange coincidence, the word is 'good'. And, let us suppose, also, that it really is equivalent – that it is, as the Oxford English Dictionary puts it, 'the most general adjective of commendation' [...] If the missionary has mastered his vocabulary, he can, so long as he uses the word evaluatively and not descriptively, communicate with them quite happily. They know that when he uses the word he is commending the person or object that he applies it to. The only thing they find odd is that he applies it to such unexpected people, people who are meek and gentle and do not collect large quantities of scalps; whereas they themselves are accustomed to commend people who are bold and burly and collect more scalps than the average.
>
> We thus have a situation which would appear paradoxical to someone who thought 'good' [...] was a quality word like 'red'. Even if the

> qualities in people which the missionary commended had nothing in common with the qualities which the cannibals commended, yet they would both mean what the word 'good' meant. If good were like 'red', this would be impossible; for the cannibals' word and the English word would not be synonymous [...] It is because in its primary evaluative meaning 'good' means neither of these things, but is in both languages the most general adjective of commendation, that the missionary can use it to teach the cannibals Christian morals. (Hare 1952 pp. 148-9)

Hare's argument, in outline, is that the missionary *means* the same thing by 'good' as the cannibal. Synonymy is preserved even across very different patterns of usage. The cannibal applies 'good' to the collecting of scalps; the missionary refrains from doing so. But, says Hare, this should not be possible if 'good' were a "quality word"—that is, has the semantic function of referring to a property such as redness. If that were the case, the cannibal would refer to a quality that the collecting of scalps instantiates. The missionary would refer to a distinct property, which is not instantiated by the collecting of scalps. The missionary and the cannibal would then not be using 'good' with the same meaning.

Is the moral realist committed to holding that 'good' is a "quality word", in Hare's sense? Certainly the realist thinks 'good' refers to the property of goodness. But whether it shares with 'red' a semantic profile which includes similar conditions under which 'good' refers to something other than goodness, is a contentious assumption.

Terry Horgan and Mark Timmons developed, in a series of papers, arguments which purport to show that all realist meta-semantic theories for moral terms will fail to adequately account for the scope of moral disagreement. This skirts the issue which faces Hare, since Horgan and Timmons do not *assume* a theory of reference for 'good' but make use of the theories of reference provided by realists themselves.

At the heart of their argument is a claim similar to that of Hare: there are possible users of moral language who intuitively disagree with each other. But the realist, they allege, will predict these communities are talking past one another, rather than disagreeing. Their case involves two communities of moral language-users, which they characterize as inhabitants of Earth and Moral Twin Earth:

> Earthlings' moral judgments and moral statements are causally regulated by some unique family of functional properties, whose essence is functionally characterizable via the generalizations of a single substantive moral theory. Suppose, too, that this theory is discoverable through moral inquiry employing coherentist methodology. For specificity, let this be some sort of consequentialist theory, which we will designate T^c.
>
> Now for Moral Twin Earth. Its inhabitants have a vocabulary that works very much like human moral vocabulary; they use the terms

> 'good' and 'bad', 'right' and 'wrong', to evaluate actions, persons, and so forth (at least those who speak twin English use these terms, whereas those who speak some other twin language use institutions, terms orthographically identical to the corresponding moral terms in the corresponding Earthly language). But on Moral Twin Earth, people's uses of twin-moral terms are causally regulated by certain natural properties distinct from those that (as we are already supposing) regulate English moral discourse. The properties tracked by twin English moral terms are also functional properties, whose essence is functionally characterizable by means of a normative moral theory. But these are non-consequentialist moral properties, whose functional essence is captured by some specific deontological theory; call this theory T^d. (Horgan and Timmons 1992: 245)

Horgan and Timmons go on to point out that, intuitively, the speakers on Earth and Moral Twin Earth should count as having important moral disagreements. While the Earthlings, as consequentialists, say 'it is ok to sacrifice one person to save many', the Twin Earthlings, as non-consequentialists, say 'it is not ok to sacrifice one person to save many'. The intuition is that, in making these utterances, the Earthlings and Twin Earthlings disagree.

There are distinct properties that fit the moral terms, as used by Earthlings and Twin Earthlings. One of these—a "consequentialist" property that is instantiated by all and only those things that promote the most good—is instantiated by the things Earthlings apply 'right' to. Another property—a "non-consequentialist" property that is instantiated only by those things actions which respect the autonomy of rational agents—is instantiated by all and only the things that Twin Earthlings apply 'right' to. Horgan and Timmons conclude (plausibly, for reasons I discuss below) that common theories of reference proposed by realists imply that Earthlings and Twin Earthlings refer to different properties with their word 'right'. Thus they fail to genuinely disagree with each other over moral matters.

This is just one type of argument from disagreement that realists have to contend with. There are other objections to realism that concern disagreement as well. These objections presuppose that disagreement exists in the first place, so grant what is being contested here, namely that realists can (and should) explain how it is that moral disagreement is possible. It is worth mentioning these objections, since they highlight the fact that even if realists can respond to Moral Twin Earth-style objections from disagreement, there are still further obstacles for the realist. One type of objection is an explanatory challenge, found in Mackie (1977). This objection alleges that, since many people have moral disagreements, the best explanation for *why* these disagreements exist is that some form if anti-realism is true. A second objection alleges that the phenomenon of disagreement coupled with realism brings with it unpalatable epistemological consequences, such as the failure of true moral beliefs to be genuine knowledge, or to have some other desirable epistemic status. Versions of this argument can be found in

Rowland (2017) and the Tersman and Risburg chapter of this volume.

2 Naturalism vs. non-naturalism

Horgan and Timmons explicitly target *naturalist* versions of moral realism with their Moral Twin Earth argument. However, challenges from disagreement will be relevant to *non-naturalist* versions of moral realism as well.

Non-naturalists must explain why it is that both speakers on Earth and Moral Twin Earth refer to the same non-natural property, and hence have a disagreement. At first glance this might seem straightforward: if goodness is the only non-natural property, then it would not be possible for Earthlings to refer to something other than non-natural goodness with their moral term 'good'. However, things are more complicated than this, for several reasons.

First, we need to explain why the only non-natural moral properties are goodness, rightness, etc. Suppose that the right actions are just those that have the best consequences without violating the autonomy of a rational agent. It is also true that there is the collection of actions which we can call the *right** actions: these are just those actions which have the best consequences, period—without regard to whether they violate the autonomy of a rational agent. The non-naturalist must the explain why there is not, in addition to the non-natural property of rightness, also non-natural property rightness*. Rightness* would better fit the Earthlings' use of 'right', so if it does exist, non-naturalist does no better at solving the Moral Twin Earth problem than naturalism does. Second, it is not enough to explain a disagreement between the Earthlings and Moral Twin Earthlings to show that there is only one (non-natural) candidate for each to refer to. The Earthlings could fail to refer to anything at all, if their use of 'right' fails to track non-natural rightness. Non-naturalists appear to have a better claim to explaining why Earthlings and Twin Earthlings do not refer to distinct properties, because they claim (perhaps without adequate support—see the first worry above) that there is a single unique non-natural property of rightness. Even if we suppose that moral terms are uniquely suited to referring to non-natural properties, a sufficient divergence between the Earthlings and Twin Earthlings will plausibly have the consequence that one of the communities is not referring to anything with their use of 'right'. Finally, just because there are non-natural moral properties, it does not follow that moral terms *must* refer to a non-natural property. If the Earthlings' use of 'right' fails to track the non-natural property of rightness, it might instead refer to a natural property that is instantiated by most of what the Earthlings call 'right'.

None of this shows that non-naturalists will inevitably fail to explain moral disagreement. However, it does show that they need to explain some significant claims, in order to provide an adequate explanation. In particular, non-naturalists will have to explain (i) why non-natural properties are especially privileged as referents for moral terms, (ii) why divergent uses of moral terms like 'right' do not result in the absence of disagreement because at least one party fails to refer to anything at all, and (iii) why non-natural properties are not sufficiently plentiful to render at least some apparent moral disagreements as verbal disputes, where the divergent communities refer to distinct non-natural properties. Non-naturalistic moral realism is not immune to challenges from disagreement.

Naturalists, on the other hand, typically face Moral Twin Earth-style challenges from moral disagreement because they explicitly commit to a naturalistic theory of reference for moral terms. Naturalistic theories of reference are modeled on theories of reference for terms that pick out natural properties. Often, these theories appear to predict that moral disagreements will not be possible between the Earthlings and Twin Earthlings, or communities like them.

For example, take the naturalistic theory of reference which inspired the original Moral Twin Earth article, from Boyd (1988). Boyd proposed a "causal regulation" theory of reference for 'good' which was modeled on a theory of reference for natural kind terms suggested by Kripke (1980) and Putnam (1975). Kripke's discussion of natural kind terms such as 'water' inspired development of the idea that 'water' refers to H_2O because the presence of H_2O is typically what causes English speakers to use their word 'water'. Importantly, the causal theory of reference implies that speakers can refer to H_2O with their world 'water' without knowing that water is H_2O, or even being aware that water has a molecular composition at all.

The advantages of a causal theory of reference are considerable for a moral realist who accepts naturalism. If goodness is a natural property, then in virtue of being natural it can causally regulate speakers' use of 'good'. Moreover, speakers do not have to know what goodness is, in order for the natural property that constitutes goodness to causally regulate their use of 'good'. Hence naturalists can explain how speakers refer to goodness, by using natural kind terms as their model. And they can explain why speakers do not need to have a complete and adequate theory of goodness, in order to successfully refer to it.

What Horgan and Timmons's Moral Twin Earth thought experiment shows is that Boyd's causal theory of reference implies that some possible users of moral language will fail to disagree. In particular, the Earthlings and Twin Earthlings are described as communities whose use of 'good' is causally regulated by different properties. The causal theory of reference implies that the Earthlings and Twin Earthlings refer to different properties, and so do not disagree when they use their word 'good' differently. Horgan and Timmons (2000) claims that a version of this argument will work for *any* naturalist theory of reference.

To summarize: Moral Twin Earth objections from disagreement have largely been leveled against naturalists. Since naturalists tie their view to a naturalistic theory of reference, such as Boyd's causal regulation theory, an objection from disagreement can be raised by constructing a case with two communities who bear the naturalistic reference-determining relation to distinct properties. But as we sketched earlier, non-naturalists may be vulnerable to similar objections.

3 Responses

Moral Twin Earth-style objections from disagreement allege that disagreement is a phenomenon that is difficult for realists to explain. The objection alleges that, while moral disagreements are apparently pervasive, realist cannot explain this. The alleged reason for this failure is that realists cannot explain why it is that communities who use their moral language differently manage to refer to the same property. I will canvass the possible avenues for realists to respond to this objection below.

Realists can respond to the challenge in a number of ways. Broadly, there are two main options. First, the realist might accept that Earthlings and Twin Earthlings refer to different properties, but argue that this is not implausible. This position needs to supply an argument that our intuitions about disagreement can be explained without appeal to co-reference. Second, realists can provide a theory of reference for moral terms which predicts that Earthlings and Twin Earthlings do not disagree. This will require supplementing or replacing the standard theories of reference that realists appeal to.

These options are not exhaustive. There are alternatives, which include denying the intuition of disagreement in the Moral Twin Earth case. Alternatively, realists who accept the intuition might hold that failing to accommodate it is not a significant strike against their view (cf. Dowell (2015)). Finally, realists can hold that while failure to accommodate the intuition of disagreement is a significant mark against their view, there are other advantages to moral realism which outweigh it. (This last strategy fits especially well within the methodological framework outlined in Enoch (2011).) I will not pursue these strategies below.

Copp (2000) provides an instance of the first main option, since he grants that the Earthlings and Twin Earthlings do not refer to the same property with their word 'good'. They do, however, disagree about what to do, since they hold incompatible views about "whether to avoid actions that are wrong or whether instead to avoid actions that are twin-wrong" (Copp 2000: 123).

More recent versions of this response focus on the *concept* of rightness, wrongness, and the like are used to make moral assertions. Plunkett and Sundell (2013: 7) argues that the disagreement between Earthlings and Twin Earthlings should be diagnosed as a dispute about which of the relevant moral concepts to use. We can use the concept the Earthlings express with 'right', which does not permit actions which violate the autonomy of rational agents. Or, we can use the concept the Twin Earthlings express with 'right' which requires actions which produce the best consequences, potentially including those that violate the autonomy of rational agents.

Of course the literal meaning of the sentences 'stealing from the rich to give to the poor is right' and 'stealing from the rich to give to the poor is not right' does not concern the concept of rightness. However, by asserting that stealing from the rich to give to the poor is 'right' or 'not right' interlocutors can be understood to be engaging in what Plunkett and Sundell call "meta-linguistic negotiation": they are implicating that we should use a particular concept of rightness. Since the Earthlings and Twin Earthlings use different concepts of rightness, they can disagree over which of those concepts *should* be used.

One issue these views face is the following. Disagreement about what to do—whether to perform a certain action, or to use a certain concept—is disagreement over a normative question. It is a matter of what one ought to do. However, similar problems involving disagreement also apply to the normative 'ought'. That is, we can imagine a community of Earthlings who use their normative 'ought' as non-consequentialists, and a community of Twin Earthlings who use their normative 'ought' as consequentialists. Intuitively, the Earthlings and Twin Earthlings will disagree in virtue of their assertions using the normative 'ought'.

The realist has trouble explaining how the Earthlings and Twin Earthlings can refer to the same thing using moral terms such as 'right' and 'good'. Prima facie, similar problems will plague an attempt to show that Earthlings and Twin Earthlings refer to the same thing in their use of the normative 'ought'. Realists who attempt to explain moral disagreement not through co-reference, but in terms of normative disagreement, have a further task ahead of them: they also have to explain how normative disagreement is possible. Many of the same issues arise here (Dunaway and McPherson 2016).

The second main strategy is to provide a theory of reference for the realist which directly explains why the Earthlings and Twin Earthlings disagree: they disagree because they both refer to the same moral properties with their terms 'right', 'good', etc., and make incompatible claims about those properties. One advantage of this strategy is that, if it is successful, it can be applied to the normative 'ought' as well. The disadvantage is that it places a substantial burden on the realist to explain why such a wide range of users of moral (and normative) terms refer to the same thing. This is a substantial disanalogy with typical natural kind terms, which appear *not* refer to the same property in suitably different environmental conditions. To illustrate, compare what Boyd's theory would say about the natural kind term 'water' in the kind of possibility described in Kripke (1980) where the clear, liquidy stuff around users of the word 'water' is not H_2O, but rather is the molecular compound XYZ. Boyd's theory would say that, since XYZ causally regulates use of 'water' in this possibility, 'water' in the mouths of these possible speakers refers to XYZ, and not H_2O. This is a plausible outcome for 'water'. But an analogous result appears not to be plausible for moral terms—as Horgan and Timmons point out, when different communities have their moral terms causally regulated by distinct properties, this does not result in a failure of the communities to refer to the same property.

Recent attempts to rescue the naturalist strategy have focused on the idea in Lewis (1983) that some properties are more eligible for reference than others. Lewis sometimes calls the properties that are more eligible for reference the "natural" properties, but this can be confusing, since they are not natural in the sense that concerns the dispute between naturalists and non-naturalists in meta-ethics. Rather, the natural properties are to be contrasted with "unnatural" properties. Highly unnatural properties are gerrymandered, such as the property grueness, and the property of being either under 6 feet tall or north of the 38th parallel. Others are not obviously gerrymandered in this way, but are still not-very-natural: the property of being a philosopher, for instance. To avoid confusion with other uses of 'natural', we can call the non-gerrymandered properties 'elite'.

The idea that elite properties are highly eligible for reference is as follows. When we use terms like 'electron' or 'cat'—terms that have a role in the physical and biological sciences—there are multiple candidate referents. There are not only cats or electrons, there are also cats-existing-before-3000 A.D., electrons-or-natural numbers, and many others. Because there are so many candidate referents of this kind, ordinary speakers will not be in a position to use their terms 'electron' or 'cat' with sufficient precision to rule out every candidate referent aside from electrons or cats. According to Lewis, the reason why electrons or cats count as the referent of 'electron' and 'cat' includes, in part, the fact that electrons and cats are more elite than the other candidate referents.

This idea has come to be known as *reference magnetism*. Reference magnetism can be extended to other descriptive terms beyond 'electron' and 'cat'. In particular, moral realists can also claim that reference magnetism applies to moral terms as well. This idea has been explored in van Roojen (2006), Edwards (2013), and Dunaway and McPherson (2016). There are different ways to implement the idea, which I will canvass below. The key idea as it applies to moral disagreement is as follows: disagreements with moral vocabulary, including the disagreement between Earthlings and Twin Earthlings, is genuine because moral properties are reference magnets. Hence, when the Earthlings apply their term 'right', without exception, to actions that have the best consequences, they don't refer to the property of having the best consequences. Instead, they refer to a nearby moral property that is more elite, and hence a reference magnet. This is the same property that the Twin Earthlings refer to. Hence there is a genuine disagreement.

There are several benefits to accepting reference magnetism for a moral realist. First, in virtue of accepting realism, realists are already committed to the metaphysical thesis that moral properties are a significant part of reality. So the claim that moral properties are highly elite, in the sense that makes them easy to refer to, will come naturally to the moral realist. Second, reference magnetism can be applied equally well to normative terms. Just as the realist view the property of rightness as highly elite, the property of all-things-considered obligation is, on the realist view, highly elite as well.

In order to give a theory which makes sense of the idea that moral rightness, or all-things-considered obligation, are highly elite, realists will have to revise some popular theses about eliteness. One such thesis is the claim that the elite properties are those that appear in our best physical theory (Lewis 1983). It is highly unlikely that our best physical theory will need to mention rightness or obligation. So, a theory that restricts the elite properties to those needed in a complete version of physics will not support the claim that rightness and obligation are reference magnets. Realists who wish to appeal to reference magnetism will have to adopt an alternative to Lewis's account of eliteness.

There are some reasons for pessimism about the possibility of formulating a theory of eliteness for the moral realist. One problem concerns the apparent eligibility of non-moral properties as reference magnets. Natural properties such as *enhancing evolutionary fitness* are highly elite, owing to their role in the science of biology. The property of enhancing evolutionary fitness also appears to fit with some possible uses of moral terms. For example, some communities inspired by Nietzsche might use their term 'right' by applying it to actions that involve the strong subjugating the weak and ruthless competition for dominance. Let us suppose that these actions also enhance the evolutionary fitness of the individuals that engage in them. The property of enhancing evolutionary fitness is pretty elite. And so, a theory of reference that includes reference magnetism appears committed to the conclusion that these communities refer to the property of enhancing evolutionary fitness.

This would be a disappointing result for the moral realist who appeals to reference magnetism to solve Moral Twin Earth-style problems from disagreement, and is developed at length Williams (2018). Enhancing evolutionary fitness is distinct from the property of moral rightness. Other communities with less severe moral views use their term 'right' to refer to moral rightness, which we are supposing is also highly elite. The problem of disagreement re-emerges.

One response to this problem for the realist is to hold that natural properties such as enhancing evolutionary fitness are in fact not candidate referents for the moral term 'right'. Reference magnetism is one factor that determines what moral terms refer to, but it needs to be applied with a less simplistic view of how moral terms are used (cf. Dunaway (2020: Ch. 4)). There are multiple properties that are part of the subject of morality. Rightness is one, but so are goodness, the property of having a reason for action, and so on. Users of moral language do more than just apply their term 'right' to the individual acts they think are right; in addition, they accept generalizations about the relationship these properties bear to each other. If they didn't do this, they would not be clearly recognizable as users of *moral* language at all. Examples include:

G The right action usually produces good outcomes.

R Typically, the right action is also the action that there is the most overall reason to perform.

Although it is possible for a community to have different views on the precise relationship between rightness, goodness, and reasons for action, communities that reject the qualified generalizations **G** and **R** by (for example) asserting 'the right action rarely produces a good outcome' would plausibly not be talking about moral rightness at all. The realist should hold that it is not only rightness that is highly elite. Goodness, being a reason, and other moral properties are likewise highly elite. Reference magnetism explains not only why 'right' refers to rightness, but also why 'good' and reason 'reason' in **G** and **R** refer to goodness and being a reason.

Consider again the claim that, in the mouths of members of the Nietzschean community, 'right' refers to the highly elite property of enhancing evolutionary fitness. Then, 'right' in **G** and **R** refers to enhancing evolutionary fitness. But if 'good' in **G** refers to goodness, and 'reason' in **R** refers the property of being a reason, then **G** and **R** are false. It is not necessary that actions which enhance evolutionary fitness by making it more likely that the actor pass on its genes generally produce good outcomes. Likewise, we don't typically have most overall reason to do what promotes evolutionary fitness. This shows that the realist who appeals to reference magnetism is not committed to the possibility of a community which speaks a language that is exactly like English except for one departure, where their 'right' refers to the property of enhancing evolutionary fitness, rather than rightness.

The problem for reference magnetism arises only if the Nietzschean community refers to the property of enhancing evolutionary fitness with their 'right' because there are *also* another highly elite property that is usually instantiated by the outcomes produced by fitness-enhancing actions. This is what is needed to make **G** in the mouths of the Nietzscheans true. Similarly, the Nietzschean community refers to the property of enhancing evolutionary fitness with their 'right' only if there is also some other highly elite property 'reason' can refer to, to make **B**, so interpreted, true. There is no guarantee that such referents can be found. (See Lewis (1984) for a response along these lines to a puzzle about reference from Putnam (1981).) The upshot is that rekindling Moral Twin Earth-style objections from disagreement against a realist theory that appeals to reference magnetism requires a substantial amount of additional work to succeed.

This solution, however, raises a second worry for the realist who appeals to reference magnetism. Even if reference magnetism provides the realist with an outline of a response to Moral Twin Earth-style objections from disagreement, the realist still appears to need an account of why it is that moral properties are highly elite. If the realist needs multiple moral properties to be elite—an assumption the above response relies on—then this explanatory task becomes even more difficult. That is, the realist needs to explain not only why rightness is elite, the realist also needs to explain why goodness and being a reason are also elite.

We have already noted that the moral realist needs to reject Lewis's claim that the highly elite properties are those that appear in a fully complete theory of physics. van Roojen (2006) appeals to different types of naturalness or eliteness, holding that nothing is elite simpliciter. Rightness, goodness, and other moral properties are elite relative to the discipline of morality, while cathood and the property of enhancing evolutionary fitness are highly elite relative to the discipline of biology. The idea here is that biological terms with refer to the properties that are highly natural relative to biology, astronomical terms will refer to properties that are highly natural relative to astronomy, and moral terms will refer to properties that are highly natural relative to morality.

An alternative to this idea is that there is a single kind of eliteness, which is not discipline-relative, but is primitive. In saying that it is primitive, we mean that no explanation can be given for why some properties are elite, and others are not. This does not necessarily mean that the moral realist has free license to simply stipulate that exactly the properties which are needed to make moral realism avoid the Moral Twin Earth problem are are elite. A mere assertion that the consequentialist property is an elite moral property would, even if it were true, not be a *knowledgeable* assertion. The primitivist conception of eliteness can be coupled with epistemic standards to ensure a disciplined, informative theory of reference for the realist. One option is to hold that mature disciplines, including biology and moral theory, give us epistemic access to which properties are primitively elite. Knowing that a mature biological or moral theory makes use of a certain property is a good way to know that the relevant biological or moral property is elite (cf. Dunaway (2020)).

Some questions regarding reference magnetism for moral terms remain to be explored. It is unclear to what extent a naturalist will be able to appeal to reference magnetism for moral terms. If there is no naturalist-friendly way to implement reference magnetism, then insofar as it is a required component of a realist-friendly response to objections from disagreement, disagreement will tell against naturalism, and in favor of non-naturalism. Further, the commitments of reference magnetism for the realist will need to be compared against the costs and benefits of alternatives to realism. We haven't canvassed the attractions of, and problems for, fictionalism, relativism, expressivism, and other alternatives to realism here. We can say that disagreement does not supply a decisive objection to realism, because the realist has several options to accommodate the phenomenon. Still, opponents of realism often complain that the view is metaphysically extravagant. The resources the realist needs to respond to arguments from disagreement will only serve to amplify these complaints.

References

Blackburn, S. (1988), 'Attitudes and Contents', *Ethics* **98**(3), 501–517.

Boyd, R. N. (1988), How to be a Moral Realist, *in* G. Sayre-McCord, ed., 'Essays on Moral Realism', Cornell University Press.

Copp, D. (2000), 'Milk, Honey, and The Good Life on Moral Twin Earth', *Synthese* **124**(1-2), 113–137.

Dowell, J. L. (2015), The Metaethical Insignificance of Moral Twin Earth, *in*

R. Shafer-Landau, ed., ‘Oxford Studies in Metaethics, vol. 11’, Oxford University Press.

Dunaway, B. (2020), *Reality and Morality*, Oxford University Press.

Dunaway, B. and McPherson, T. (2016), ‘Reference Magnetism as a Solution to the Moral Twin Earth Problem’, *Ergo* **3**(25), 639–679.

Edwards, D. (2013), ‘The Eligibility of Ethical Naturalism’, *Pacific Philosophical Quarterly* **94**(1), 1–18.

Enoch, D. (2011), *Taking Morality Seriously: A Defense of Robust Realism*, Oxford University Press.

Gibbard, A. (2003), *Thinking How to Live*, Harvard University Press.

Horgan, T. and Timmons, M. (1992), ‘Troubles on Moral Twin Earth: Moral Queerness Revived’, *Synthese* **92**(2), 221–260.

Horgan, T. and Timmons, M. (2000), ‘Copping Out on Moral Twin Earth’, *Synthese* **124**(1-2), 139–152.

Kripke, S. (1980), *Naming and Necessity*, Harvard University Press.

Lewis, D. (1983), ‘New Work for a Theory of Universals’, *Australasian Journal of Philosophy* **61**(4), 343–377.

Lewis, D. (1984), ‘Putnam’s Paradox’, *Australasian Journal of Philosophy* **62**(3), 221–236.

Mackie, J. (1977), *Ethics: Inventing Right and Wrong*, Penguin.

Plunkett, D. and Sundell, T. (2013), ‘Disagreement and the semantics of normative disagreemeent and the semantics of normative and evaluative terms’, *Philosophers’ Imprint* **13**(23), 1–37.

Putnam, H. (1975), The Meaning of ’Meaning’, *in* ‘Mind, Language, and Reality: Philosophical Papers, vol. 2’, Cambridge University Press, pp. 215–271.

Putnam, H. (1981), *Reason, Truth and History*, Cambridge Univsersity Press.

Rowland, R. (2017), ‘The Significance of Fundamental Moral Disagreement’, *Noûs* **51**(4), 802–831.

van Roojen, M. (2006), Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument, *in* R. Shafer-Landau, ed., ‘Oxford Studies in Metaethics, vol. 1’, Oxford University Press.

Williams, J. R. G. (2018), ‘Normative Reference Magnets’, *The Philosophical Review* **127**(1), 41–71.